Lydia internal examiner comments

Andrew Simpkin

2024-02-02

## Overall typographical comments

* US/UK spelling, check modelled vs modeled, normalized v normalised - e.g. do a search for “ize” and replace with “ise” where applicable. Check Figure legends.
* Formatting chapter headings: “1 Chapter 1” not good.
* Acronyms: define first then use forever. Lots of jumping back to long form, e.g. Kaplan-Meier then KM then Kaplan-Meier again.
* Acronyms in table and figure captions: need to be explained, this is mainly for Chapter 5 AI and AD models.
* Punctuation: overuse of commas instead of brackets or colons [page 54 great example].
* Signposting: at the start (and maybe end) or each Chapter.
* Signposting: better use of subsections to describe methods before results - think in terms of reproducibility.
* Figure and Tables are usually numbered by Chapter rather than a running total - this might be helpful but not necessary.

## Overall statistical reporting comments

* Using p<0.05 instead of reporting exact p-value is not recommended.
* Be consistent with decimal places: sometimes one, two or three.

## Chapter 1

* You discuss the idea of improving predictive models - this doesn’t carry through later as we don’t compare prediction performance with/without CNA. Is this something for future work?
* Clarify if Figures (e.g. Fig 1, 2) are original or from another publication in the caption.

## Chapter 2

* Need more detail on predictive accuracy of previous methods - how accurate are CES, CIN, CAAI for example? Can you provide simple accuracy/AUC statistics for the key methods?
* You say missing data are “unlikely to invoke bias” - what bias is being avoided? Clarify in text.
* Could missing data be imputed? Add a reason why this has not been done here.
* If p>0.05 does this mean two groups are similar? Change to say “no significant difference”.
* What is new or novel in this chapter? Perhaps put a summary of this at the end.

## Chapter 3

* What are the measures of node impurity for categorical data - Gini info and entropy - can you briefly explain.
* Use past tense in describing results. Analysis is done in steps and present tense does not work well.
* Provide HR and CI from Cox models rather than just p-values.
* What is the depth of a tree (page 66 depth of 4)?
* Talk through model building in Section 3.2 … univariate models used to find significant variabless, then dropped “similar variables”, then CNA score still significant. But why drop anything if prediction is the goal, and why use significance if inference is the goal? Needs more detail, perhaps in a subsection?

## Chapter 4

* How were the trees for DGEA chosen (end of page 123)? Clarify.
* Is this model run separately for each gene? Clarify.
* Discuss comparative results - should more/less overlap be expected?
* Can gene expression be used to improve predictive ability? Future work?

## Chapter 5

* Why curtail number of CNA to amp, neut, del? Changepoint models could be applied to CNA counts. Future work?
* AINM models do not use a random intercept for individuals, has this been tested? Discuss.
* Need to walk through the simulation study - are the estimates correct - what is the truth? 0=0 is fine, but the others with values are highly variable … is the mean where it’s supposed to be?

## Chapter 6

* KM analysis: are changepoints good or bad? Conflicting results in presentation. Discuss.
* Could changepoints and CNA Scores be used in combination to predict survival? Discuss.

## Chapter 7

* Can you use GE + CNA score for prediction?
* Can you use combine parts 1 + 2 for prediction?
* What about allowing copy number > 2 in changepoint detection?

## Bibliography

* Check month before year, mostly abbreviated but some numerical months.
* Check journal names - all should be abbreviated.

## Appendices

* Add a short description at the start of each Appendix detailing what it contains.